ISSN 1000-0054 清华大学学报 (自然科学版) 2006 年 第 46 卷 第 1 期 23/41
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2006, Vol. 46, No. 1 86-89

基于语音声学特征的情感信息识别

蒋丹宁, 蔡莲红

(清华大学 计算机科学与技术系, 北京 100084)

摘 要: 为提高情感语音识别的正确率, 研究了声学参数的统计特征和时序特征在区分情感中的作用, 并提出了一种将两者相融合的情感识别方法。在提取出基本的韵律参数和频谱参数后, 首先利用 PNN (probabilistic neural network) 和 HMM (hidden markov model) 分别对声学参数的统计特征和时序特征进行处理。计算它们各自属于每类情感的概率, 获得采用加法规则和乘法规则融合统计特征和时序特征的识别结果。实验结果表明: 各组特征在区分情感方面的侧重不尽相同, 通过特征融合, 平均识别正确率相较单独采用统计特征或时序特征均有提高, 在最好情况下达到了 92.9%。这说明了该方法的有效性。

关键词: 语言识别; 模式识别; 情感信息处理; 声学特征

中图分类号: TP 391.4 **文献标识码**: A

文章编号: 1000-0054(2006)01-0086-04

Speech emotion recognition using acoustic features

JIANG Danning, CAI Lianhong

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: A speech emotion recognition algorithm was developed based on the statistical and temporal features of the acoustic parameters for discriminating between emotions. The system first extracted the basic prosody parameters and spectral parameters, then used a PNN (probabilistic neural network) to model the statistic features and a HMM (hidden Markov model) to model the temporal features. The sum and product rules were used to combine the probabilities from each group of features for the final decision. Experiments on the Chinese speech corpus showed how the statistical and temporal features tend to reflect different aspects of emotions. The accuracy rate obtained by feature combination is higher than that by each group alone, reaching a maximum of 92.9%.

Key words: speech recognition; pattern recognition; emotion information processing; acoustic features

大量的研究显示, 语音的情感信息包含在多种声学参数的变化中, 文[1]将其归纳为基频、时长、能量和频谱 4 个方面。在主要的情感类别中, 愤怒和高兴均表现为基频均值、变化范围和方差的提高, 能量的加强, 以及频谱中高频成分的增加。与此相反, 悲伤对应于基频均值和变化范围的减小, 能量的减弱, 语速的减慢, 以及频谱中高频成分的降低。害怕的特征除了基频均值、变化范围和频谱中高频成分的增加外, 还包括基频曲线上抖动的加强和语速的加快。惊讶则表现为很宽的基频变化范围, 以及稍减慢的语速。声学参数随时间的变化曲线也负载了一定的情感信息。例如, 文[2]的研究发现, 愤怒与高兴相比, 基频曲线在句末的下倾更为剧烈。

一般地, 目前基于声学特征的情感识别仅单独采用了声学参数的统计特征或时序特征[3]。前者计算参数的各个统计值并组成一个具有固定维数的特征向量, 识别模型是针对固定维数特征的, 如 GMM (gaussian mixture model), ANN (artificial neural network), SVM (support vector machine) 等; 后者通常采用 HMM (hidden markov model) 对特征的时间序列进行处理。由于统计特征和时序特征可能在区分情感方面各有侧重[1], 因此为进一步提高识别性能, 需要将它们相融合。一种融合方法是在对时序特征进行长度归一之后, 与统计特征一同输入到处理固定维数特征向量的模型中[4], 但问题是很难找到一种令人信服的长度归一方式。本文提出了一种能够融合统计特征和时序特征的情感识别方法。

1 特征提取

1.1 基本的声学参数

所采用的声学参数包括韵律参数和频谱参数。在韵律参数中, 由于能量的提取受到录音时音量设

收稿日期: 2005-01-12
基金项目: 国家自然科学基金资助项目 (60433030, 60418012)
作者简介: 蒋丹宁 (1979-), 女 (汉), 辽宁, 博士研究生。
通讯联系人: 蔡莲红, 教授, E-mail: clh-dcs@tsinghua.edu.cn

置的影响，因此仅提取了基频和时长参数。基频由一种修正后的自相关算法 YIN[5]提取，参数估计的间隔为 2ms。时长参数通过标注工具 VisualSpeech 获得。该工具能够自动估计音节边界，并提供了手工修改的界面。在提取出基频曲线之后，同时计算了它的一阶和二阶差分曲线。

频谱参数的提取借鉴了音频和音乐分类方面的相关研究[6,7]。所提取的频谱参数分别反映了能量在不同频段的分布情况，频谱在局部时间段的变化特性，以及信号周期性的好坏。具体包括：

1) Mel 频率倒谱系数 MFCC，度量了语音频谱的能量包络。由于 Mel 频率符合人类的听觉特性，同时倒谱运算具有良好的解卷性能，因此 MFCC 参数被广泛地应用在语音识别、说话人识别、音频和音乐分类中。

2) 频谱质心参数 S_c，即以各频率幅度为加权系数的频率中心，计算公式为

$$S_c = \frac{\sum_{n=1}^{N} nA(n)}{\sum_{n=1}^{N} A(n)}, \quad (1)$$

其中 $A(n)$ 为第 n 条谱线所对应的幅度。

3) 频谱截止参数 S_r，即频谱能量从低频积累到占总能量 85% 时的频率点，反映了高频能量衰减的程度，即

$$\sum_{n=1}^{S_r} A(n) = 0.85 \sum_{n=1}^{N} A(n). \quad (2)$$

4) 频谱变迁参数 S_f，即相邻两帧频谱之间的距离，计算公式为

$$S_f = \sum_{n=1}^{N} [A_i(n) - A_{i-1}(n)]^2, \quad (3)$$

其中 $A_i(n)$、$A_{i-1}(n)$ 分别为当前帧和前一帧的幅度谱。该参数反映了频谱在局部时间段中的变化特性。

5) 频带周期性参数 S_p。该参数反映了各频带信号的周期性好坏。在计算时，首先将语音信号通过若干个具有不同频率范围的滤波器，分别为 0～500Hz，500～1 000Hz，1 000～2 000Hz，2 000～4 000Hz。对于通过第 j 个频带的信号，在当前帧和前一帧的范围内计算归一化的相关函数

$$R_j(k) = \frac{\sum_{m=1}^{M} s_j(m-k)s_j(m)}{\sqrt{\sum_{m=1}^{M} s_j^2(m-k)}\sqrt{\sum_{m=1}^{M} s_j^2(m)}}. \quad (4)$$

其中：$s_j(m)$ 为通过相应频带的信号，若 $m > 0$ 则 $s_j(m)$ 包含在当前帧；否则在前一帧的范围内。若信号具有很好的周期性，则 $R_j(k)$ 应在对应于基音周期整倍数的位置上出现明显的峰；相反若信号为噪音，则 $R_j(k)$ 应是平坦的。定义第 j 个频带信号的周期性参数为 $R_j(k)$ 中最大峰值（$k=0$ 处除外）。选取 4 个频带周期性参数的平均值度量语音信号的周期性。虽然 S_p 参数的计算是在时域信号上进行的，但由于它所反映的信号周期性对频谱分布有直接影响，因此也被归为频谱参数。

所有的频谱参数仅在浊音段提取，分析帧长为 25ms，帧移为 12.5ms。在提取出各个频谱参数之后，同时也计算了它们的一阶差分。

1.2　统计特征

在提取了基本的声学参数之后，它们的统计特征为参数在句子范围内的各种统计值。韵律统计特征包括基频及其一阶、二阶差分参数的平均值、标准差、最大值和变化范围，以及音节时长的平均值和标准差。频谱统计特征包括各个频谱参数及其一阶差分的平均值和标准差。

1.3　时序特征

在提取时序特征之前，为了消除统计特征的影响以便研究两者各自独立的作用，首先将各个参数曲线根据它们在句子范围内的平均值和标准差进行归一处理。虽然，一般针对西方语言的工作直接将参数曲线作为时序特征[3]，但由于汉语是一种声调语言，句子的基频曲线与每个音节的声调曲线密切相关，HMM 模型很容易受到局部曲线形状的影响而得不到好的效果。因此，本文的韵律时序特征是基于音节计算的。特征向量序列的长度等于句子的音节总数，其中的每个特征向量为韵律参数在相应音节范围内的统计值，包括基频及其一阶、二阶差分的平均值、标准差、最大值和变化范围，以及该音节的时长。对于频谱参数，直接将归一后的各个参数及其一阶差分曲线作为时序特征。

2　识别模型

2.1　针对统计特征的识别模型

本文采用 PNN 模型对统计特征进行识别。PNN 的网络结构为 3 层。其中第 1 层节点的传递函数为高斯型函数，个数等于训练集中的样本个数，每个节点的权值向量都分别等于训练集中某个样本的输入向量。在识别过程中，第 1 层网络的作用是计算输入特征与各节点权值向量之间的相似程度。第 2

层的每个节点分别代表 1 类情感, 其传递函数为线性函数。当第 1 层节点所表示的特征向量属于第 2 层节点所表示的情感类别时, 相应的权值被设为 1, 否则为 0。则第 2 层的作用是将第 1 层计算出的与训练集中每个特征向量之间的距离加权相加, 作为输入特征向量属于每类情感的概率输出到第 3 层。第 3 层(输出层) 节点的传递函数为竞争型的传递函数, 它从第 2 层节点中选择 1 个输出概率最大的作为最终的识别结果。PNN 不需要复杂的训练过程, 它主要通过存储训练集中的样本进行识别。当训练集包含足够多的样本时, PNN 具有很好的性能。

2.2 针对时序特征的识别模型

时序特征由 HMM 模型进行处理。本文所采用的模型为前向连接模型, 具有 4 个状态, 每个状态的概率分布由包含 4 个 Gauss 成分的 GMM 模拟。

2.3 融合统计特征和时序特征的识别

设从样本 x 中提取出 N 组特征(包括统计特征和时序特征), 记为 $f_1 \sim f_N$, 它们通过 PNN 或 HMM 模型得到的属于第 i 类情感的概率为 $P(C_i|f_1) \sim P(C_i|f_N)$, 则最终的识别结果 r 为

$$r = \arg\max_i F[P(C_i|f_1), \cdots, P(C_i|f_N)], \quad (5)$$

其中函数 F 表示融合规则。本文尝试了两种不同的融合规则[8]。

1) 乘法规则。即

$$F[P(C_i|f_1), \cdots, P(C_i|f_N)] = \prod_{j=1}^{N} P(C_i|f_j). \quad (6)$$

2) 加法规则。即

$$F[P(C_i|f_1), \cdots, P(C_i|f_N)] = \sum_{j=1}^{N} P(C_i|f_j). \quad (7)$$

3 实验语料

本文所采用的汉语情感语料包含愤怒、害怕、高兴、悲伤、惊讶、中性 6 类, 每类情感包括约 200 句语句。为了方便情感表达, 各类情感语音的文本并不相同, 但它们均包含了不同的句子类型(陈述句和疑问句), 句子长度, 以及声调和重音分布等情况。所有的情感语句均由一名女性发音人在安静环境下录音得到, 并保存为 16 kHz 采样率, 16 bit 量化, 单声道的波形文件。

4 识别实验

在识别中采用了交叉检验技术。所有语句被平均分为 5 份, 而识别实验也相应地进行 5 次, 每次分别将其中的 1 份数据作为测试集, 其余的 4 份作为训练集。取 5 次实验的平均值作为识别结果。采用交叉检验技术能够降低随机因素的影响, 提高识别结果的可信度。

表 1 比较了仅采用声学参数的统计特征或时序特征, 以及将统计特征和时序特征相融合后的平均正确率。在同时采用韵律参数和频谱参数的时序特征时, 由于两者的时间粒度不同而无法合并为同一个特征向量序列, 因此表中的相应结果是将它们各自处理之后, 通过加法规则融合得到的(乘法规则的融合结果是 87.6%)。表 1 表明, 在所有情况下, 融合统计特征和时序特征能够得到优于单独使用统计特征或时序特征的情感识别性能。在采用加法规则融合所有四组特征(韵律统计, 韵律时序, 频谱统计, 频谱时序)时, 平均正确率达到了 92.9%。

表 1 融合统计特征和时序特征前后的分类正确率

%

声学参数	统计特征	时序特征	乘法规则	加法规则
韵律参数	83.1	60.5	87.7	86.7
频谱参数	86.5	87.6	89.1	88.5
韵律+频谱	90.2	88.0	91.1	92.9

为更清楚地说明各类情感之间的混淆情况, 定义第 i 类情感 C_i 和第 j 类情感 C_j 之间的混淆度 I_{ij} 为

$$I_{ij} = \frac{P(r=i|x \in C_j) + P(r=j|x \in C_i)}{2}, \quad (8)$$

其中 r 为测试样本 x 所对应的分类结果。表 2 列出了在单独采用每组特征, 以及采用加法规则融合所有四组特征时, 每两类情感之间的混淆度。

单独采用韵律统计特征时, 在高兴—惊讶和愤怒—高兴上的混淆度较大; 单独采用频谱统计特征或频谱时序特征时, 在高兴—惊讶和愤怒—惊讶上的混淆度较大。在情感理论中, 情感空间通常被划分为活动性(如愤怒是高活动性的情感, 悲伤是低活动性的情感)和评价性(如高兴是正性的情感, 愤怒是负性的情感)两个维度。参照文[1], 愤怒、高兴和惊讶 3 类情感均属于高活动性的情感。因此可推测, 韵律统计特征、频谱统计特征和频谱时序特征倾向于在活动性的维度上区分情感, 而在活动性相近的愤怒、高兴和惊讶之间的区分性较差。与此不同的是, 韵律时序特征在活动性有较明显差别的悲伤—中

性、惊讶—中性、高兴—中性上的混淆度较大。这说明韵律时序特征不能有效地在活动性的维度上区分情感，但它在活动性相近而评价性不同的愤怒—高兴、愤怒—惊讶上的区分性相对较好。

表 2　融合统计特征和时序特征前后的混淆度

%

混淆度	韵律统计	频谱统计	韵律时序	频谱时序	加法规则
高兴—惊讶	17.6+	16.4+	23.5+	21.0+	14.0+
愤怒—害怕	5.5	0.3	11.2+	0.0	0.0
愤怒—高兴	15.9+	3.3	4.5	4.7	1.9
愤怒—惊讶	5.1	10.2+	4.4	8.1+	2.8
害怕—高兴	0.0	0.5	7.4	0.0	0.3
害怕—惊讶	0.0	0.0	7.5	0.0	0.0
愤怒—悲伤	0.0	0.5	0.0	0.0	0.5
害怕—悲伤	0.8	1.7	6.1	0.0	0.0
悲伤—中性	1.5	0.0	11.9+	1.0	0.5
惊讶—中性	0.0	0.0	10.1+	0.0	0.0
高兴—中性	1.1	0.3	15.9+	0.0	0.3
害怕—中性	1.2	7.5	7.4	3.4	0.5
愤怒—中性	1.9	0.0	2.9	0.0	0.0
高兴—悲伤	0.0	0.5	2.2	0.0	0.0
悲伤—惊讶	0.0	0.0	3.5	0.0	0.0

注：上标“+”表示相应的混淆度大于 8.0%。

几乎在每两类情感之间，特征融合之后的混淆度均明显小于单独采用其中任何一组特征时混淆度的最小值。除高兴—惊讶之外，所有的情感对均能够被有效区分。

对于高兴—惊讶上混淆度较高的情况，一方面是因为这两类情感在声学上相对接近，另一方面也是因为声学特征同时受到了其他表意相关因素的影响。例如，汉语作为声调语言，音节存在四声的变化，影响了基频的平均音高和变化范围；重读的音节具有较高的基频和较长的时长；在韵律结构的边界处常常会出现一些特定的声学现象；疑问句相对陈述句具有较高的平均基频和上扬的基频曲线；等等。这些因素影响了每类情感中声学特征的分布，加剧了情感类别之间的混淆。

5　结　论

本文提出了一种能够融合声学参数的统计特征和时序特征的情感识别算法。研究表明，统计特征和时序特征在区分情感方面的侧重不尽相同。统计特征能够较好地区分激发度不同的情感，而时序特征则倾向于区分激发度相似但评价性不同的情感。在融合了韵律参数及频谱参数的统计特征和时序特征之后，除高兴—惊讶之外，每两类情感之间的混淆度均有了明显的降低，最高的平均正确率达到了 92.9%。这说明了本文所提出识别方法的有效性。

参考文献　(References)

[1] Cowie R, Cowie E D, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction [J]. *IEEE Signal Processing Magazine*, 2001, **18**(1): 32－80.

[2] Paeschke A, Sendlmeier W F. Prosodic characteristics of emotional speech: measurements of fundamental frequency movements [A]. Proc of ISCA Workshop on Speech and Emotion [C]. Northern Ireland: Textflow, 2000. 75－80.

[3] Schuller B, Rigoll G, Lang M. Hidden markov model-based speech emotion recognition [A]. Proc of ICASSP'03 [C]. New York: IEEE Press, 2003. II, 1－4.

[4] 赵力，蒋春晖，邹采荣，等. 语音信号中的情感特征分析和识别的研究 [J]. 电子学报，2004, **32**(4): 606－609.
ZHAO Li, JIANG Chunhui, ZOU Cairong, et al. A study on emotional feature analysis and recognition in speech [J]. *ACTA ELECTRONICA SINICA*, 2004, **32**(4): 606－609. (in Chinese)

[5] Cheveign A D, Kawahara H. YIN: A fundamental frequency estimator for speech and music [J]. *J Acoust Soc Am*, 2002, **111**(4): 1917－1930.

[6] Tzanetakis G, Cook P. Musical genre classification of audio signals [J]. *IEEE Transactions on Speech and Audio Processing*, 2002, **10**(5): 293－302.

[7] Lu L, Zhang H J, Jiang H. Content analysis of audio classification and segmentation [J]. *IEEE Transactions on Speech and Audio Processing*, 2002, **10**(7): 504－516.

[8] Kittler J, Hatef M, Duin R P, et al. On combining classifiers [J]. *IEEE Transactions on Pattern Analysis and Machine Learning*, 1998, **20**(3): 226－239.